# **Web Archiving of Preprint Literature**



Project Title	Web Archiving of Preprint Literature
Project Slimmary	To capture and collect VA literature residing in various preprint repositories for discovery and preservation.
Country	United States

## **Project Description**

- Investigate various Preprint Severs platforms (e.g. OSF Preprints, etc.),
- Investigate web archiving data formats e.g. Web ARChive (WARC), WARC Encapsulated Text (WET)/Parsed Text, Internet Archive, Googlebot, Heritrix, and other open source platforms, etc.
- Investigate Joint Information Systems Committee (JISC) Sherpa Services (e.g. Sherpa Romeo, Sherpa Juliet, OpenDOAR, etc.).
- Learn various standards and authorities; International Organization for Standardization (ISO), National Information Standards Organization (NISO), etc.
- Develop prototype (i.e. crawler) to source VA literature metadata via batch or harvest process.

#### **Required Skills or Interests**

Skill(s)
Data analysis
Data visualization
Design thinking
Software development
Website design

### **Additional Information**

Desired Skills: Data entry Microsoft Office (Access, Excel, Word, Publisher, SharePoint, etc.)
GitHub or Open Source Software Project Development experience.
HTML
XML
XHTML
JAVA
JSON
Python
Structured Query Language (SQL)
RDF

# **Language Requirements**

None